Sequence Diversity within STR Loci: Applications to Human Identification

Peter M. Vallone, Ph.D.
Leader, Applied Genetics Group

10th SFAF Meeting
Santa Fe, New Mexico
May 29, 2015

National Institute of Standards and Technology
U.S. Department of Commerce

Disclaimer

This presentation will mention commercial products, but we are in no way attempting to endorse any specific products.

NIST Disclaimer: Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

Information presented does not necessarily represent the official position of the National Institute of Standards and Technology or the U.S. Department of Justice.

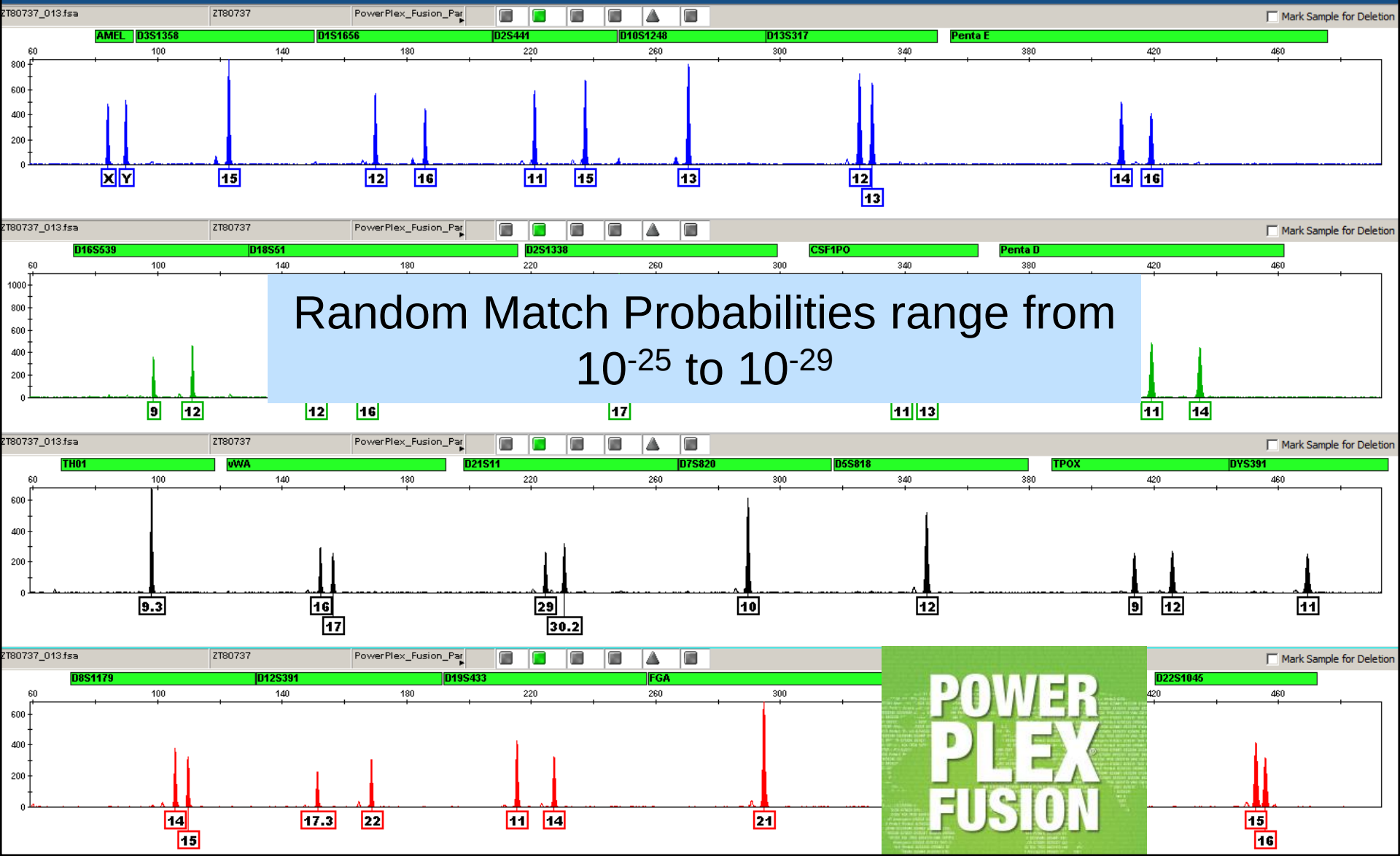
Our group receives or has received funding from the FBI Laboratory and the National Institute of Justice.

Starting with conclusions…

- 24 loci were sequenced for a set of 183 samples (22 autosomal STRs, 1 Y STR, and Amelogenin)
- Length and sequence variation within the STR region and surrounding flanking regions were detected by NGS and compared to length-based genotypes (CE)
- Loci with compound and complex repeat motifs contained the majority of ‘additional’ information by NGS
- Isoalleles and flanking region polymorphisms detected by NGS *may* assist in resolving mixtures…

Forensic STR Sequence Diversity

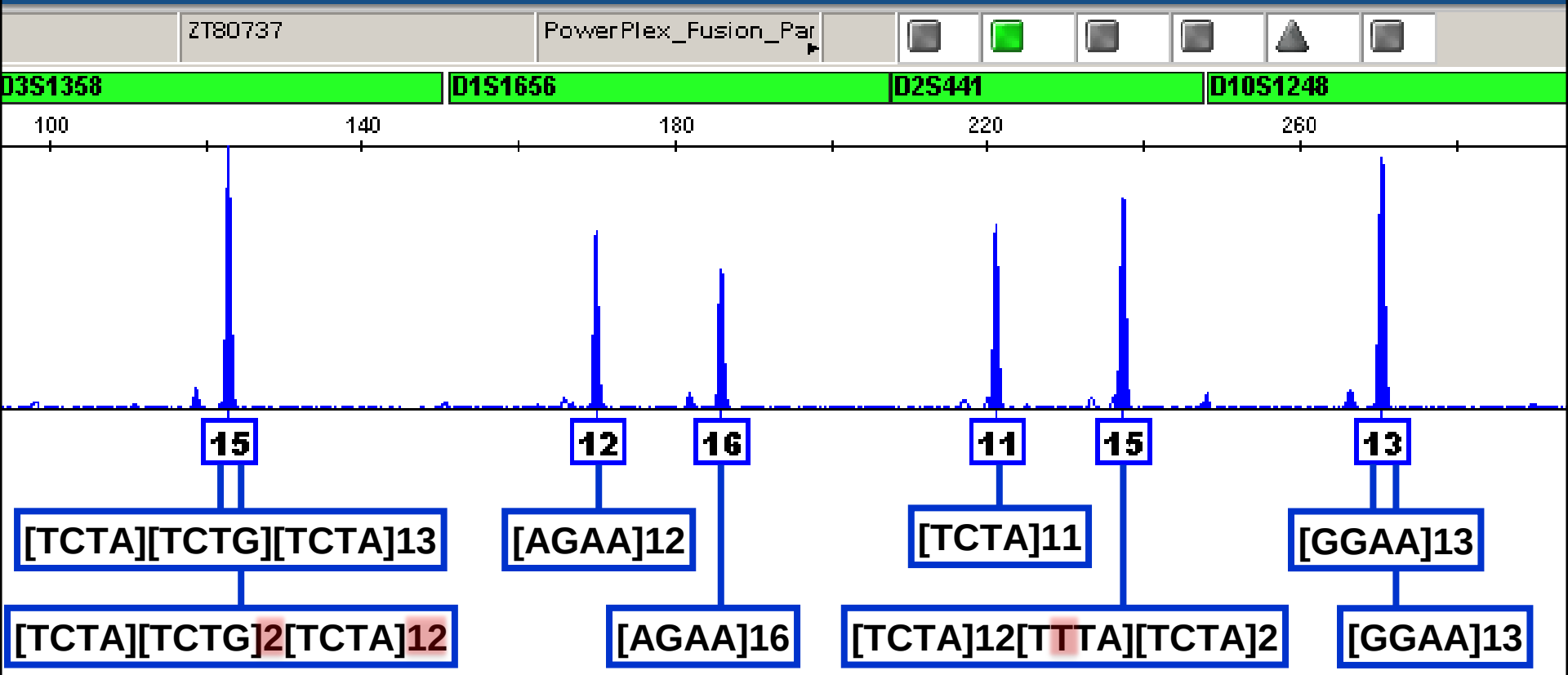
Length-based genotyping by CE methods



Random Match Probabilities range from 10^{-25} to 10^{-29}

POWER PLEX FUSION

Forensic STR Sequence Diversity



Sequence-Based Heterozygote: A locus that appears homozygous in length-based measurements (such as CE), but is heterozygous by sequence

NGS has potential for finer resolution of STR amplicons not detectable by CE-length based methods

- Additional STR alleles
- Flanking region SNPs and InDels

Applications to DNA mixtures

- Resolve isoalleles (identical by length alleles)
- Resolve minor contributor peaks from stutter

General gains of
larger multiplexing and SNP detection

Forensic STR Sequence Diversity

Methods

NIST Population Samples

- N=183
  - Caucasian (70)
  - Hispanic (45)
  - African American (68)

Amplification & Library Prep

- 1 ng input DNA
- PowerSeq Auto System (Promega)
- Illumina TruSeq HS PCR-Free

Sequencing

- MiSeq

Bioinformatics

- Sequence
- CE concordance

Pop Gen

- P_I / HET
- Repeat Region
- Flank Analysis
- Stutter

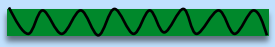
Recognition Site-Based Informatics for STRs

Computer Returns:

The length between the recognition sequences = 36 bp

Reference table in software returns a “9” allele

The sequence between the recognition sites


PCR Primers


STR Repeat Region


Recognition site (~10 nt)

[1] *http://battelleexactid.org/*

[2] STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. Warshauer et al., Forensic Sci Int Genet. 2013 (7):409-17

[3] STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. Warshauer et al., Forensic Sci Int Genet. 2015 (14):182-6

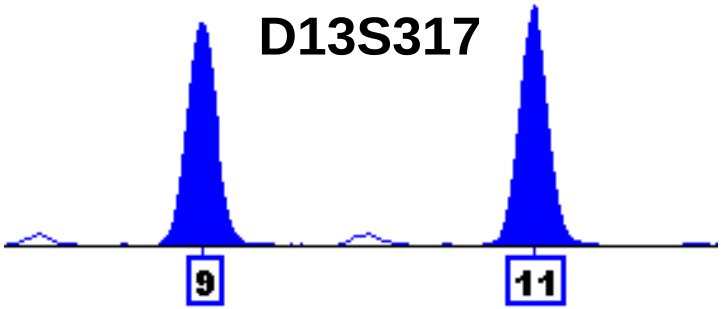
Forensic STR Sequence Diversity

CE (length-based genotype) concordance check results:

24 loci x 183 samples = 4392 loci evaluated

> 99% concordance with CE data

Forensic STR Sequence Diversity



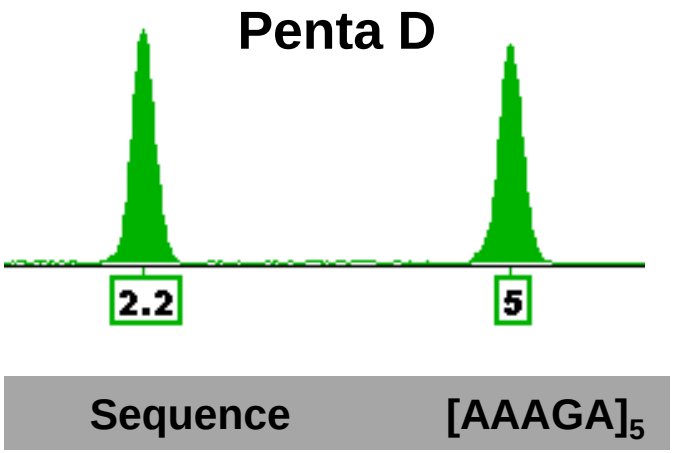
Sequence $[TATC]_{10}$,$[TATC]_{11}$

← Repeat Region NGS Recognition Region 4 bp Deletion CE Primer Binding Site→

```
TATC TATC TATC AATCAATCATCTATCTATCTTTCTGTC----TTTTTGGGCTGCCTATGGCTCAA
TATC TATC TATC AATCAATCATCTATCTATCTTTCTGTCTGTCTTTTTGGGCTGCCTATGGCTCAA
```

Flanking region InDel: Bioinformatic pipelines may reduce the region used for genotyping, resulting in deletions not being “counted” as they would via CE

Forensic STR Sequence Diversity



13 bp deletion in Recognition Region

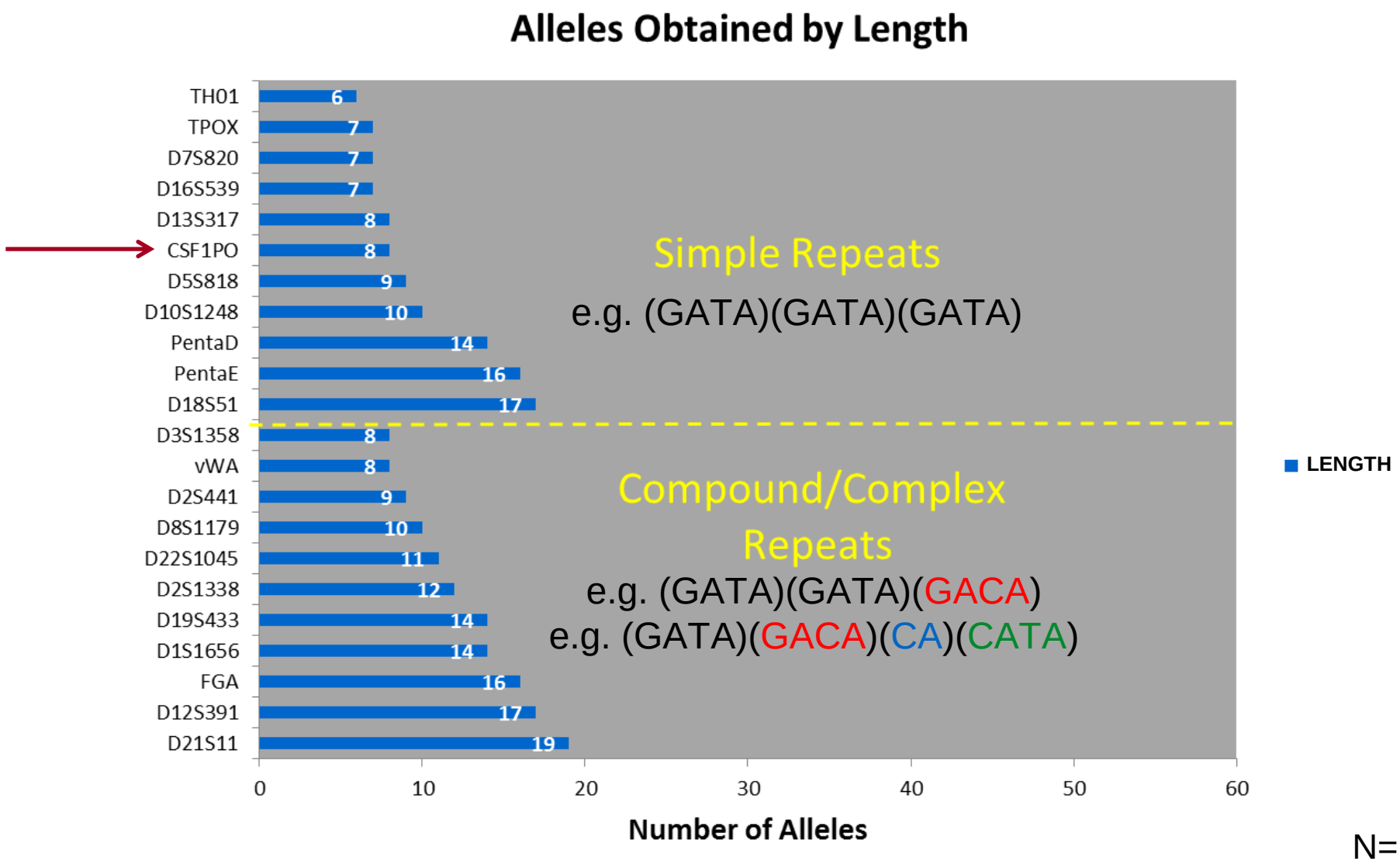
←CE Primer Binding Site

Repeat Region →

```
TAGGTTACAGAGCAAGACACCATCTCAG------------ AAAGA AAAGA AAAGA AAAGA AAA
TAGGTGACAGAGCAAGACACCATCTCAGAAAGAAAAAAAG AAAGA AAAGA AAAGA AAAGA AAA
```

Bioinformatic Null Allele: A true allele that is present within the raw sequence data but is not detected by the bioinformatic pipeline

Forensic STR Sequence Diversity



N=183

Forensic STR Sequence Diversity

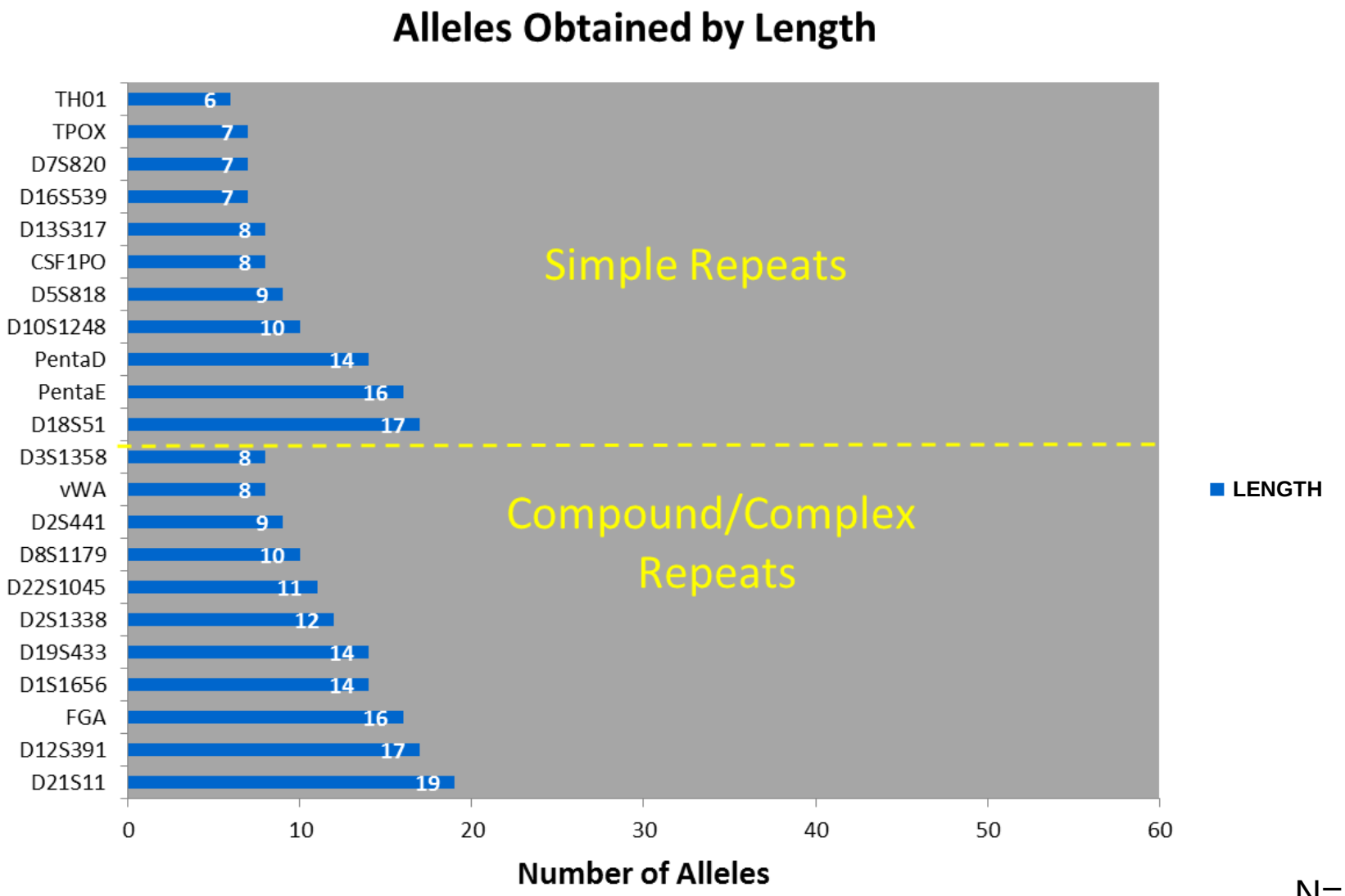
Additional Alleles by Sequence

		CSF1PO
7	[AGAT]7	AGAT AGAT AGAT AGAT AGAT AGAT AGAT
8	[AGAT]8	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
9	[AGAT]9	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
10	[AGAT]10	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
10	[AGGT][AGAT]9	AG**G**T AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
11	[AGAT]11	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
11	[AGAT]3 AGGT [AGAT]7	AGAT AGAT AGAT AG**G**T AGAT AGAT AGAT AGAT AGAT AGAT AGAT
12	[AGAT]12	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
13	[AGAT]13	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
14	[AGAT]14	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT

8 alleles by length → 10 alleles by sequence

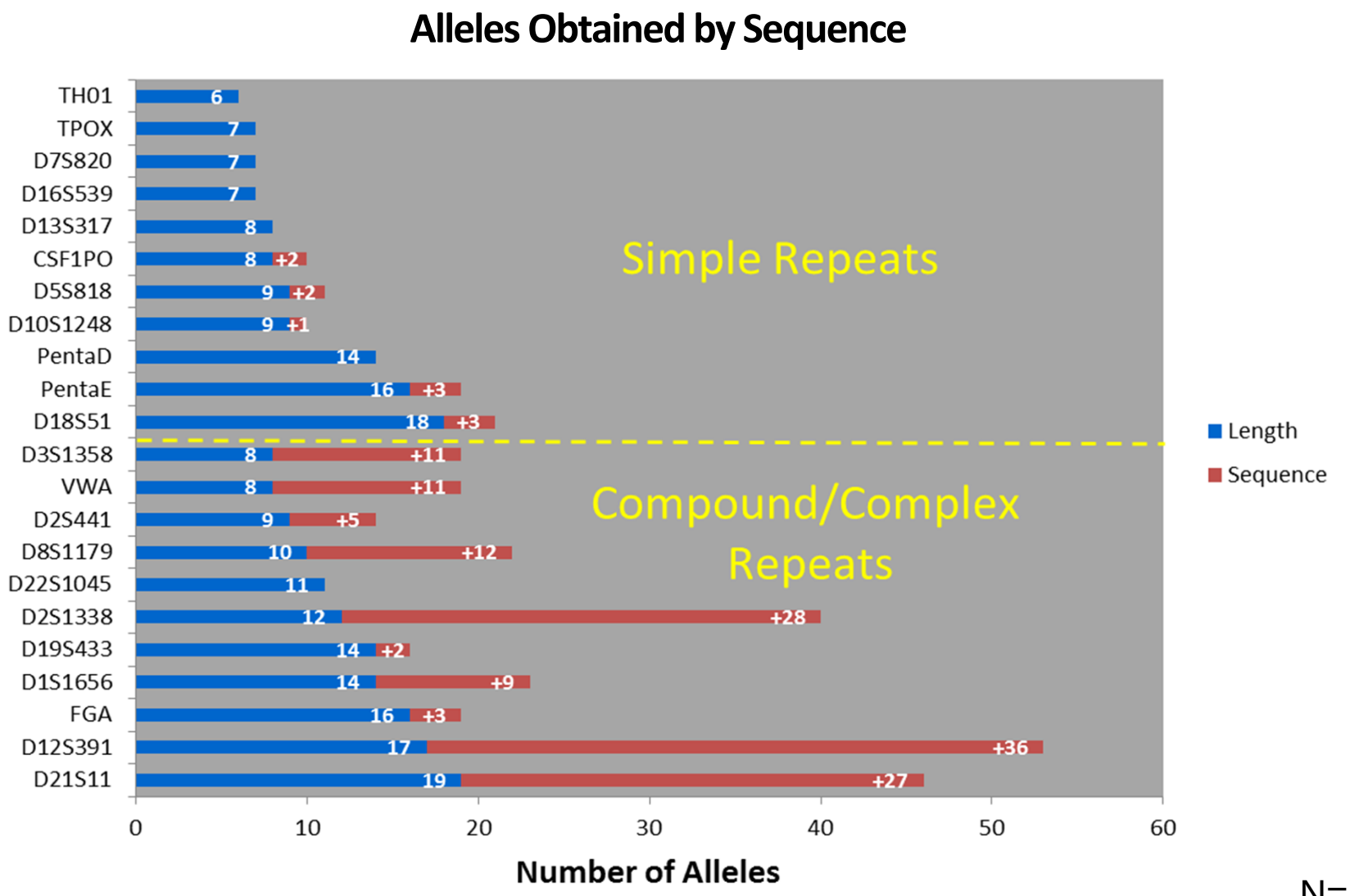
N=183

Forensic STR Sequence Diversity



N=183

Forensic STR Sequence Diversity



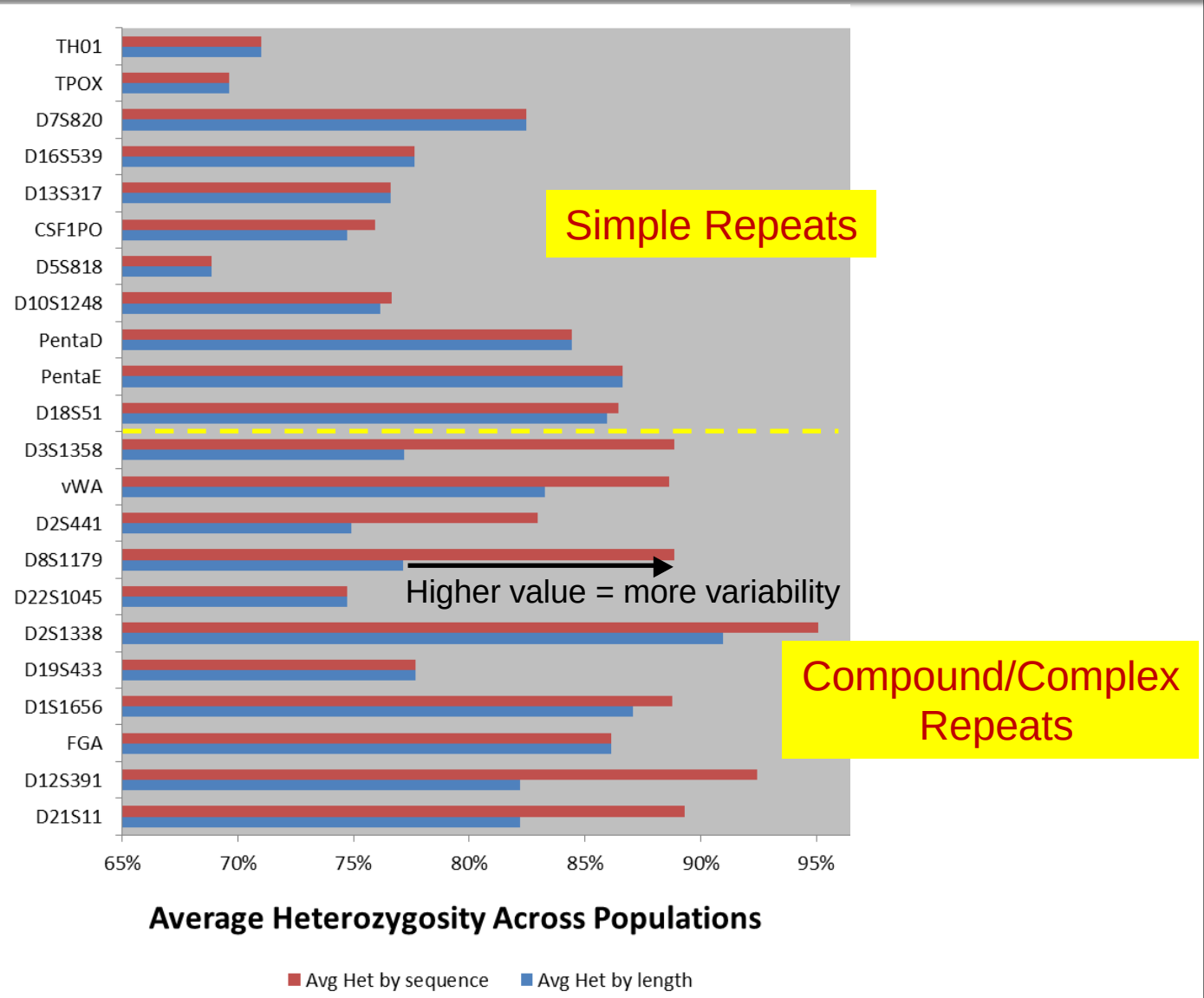
N=183

Forensic STR Sequence Diversity

Heterozygosity

$$\frac{\text{\# heterozygotes observed}}{\text{\# of loci tested}}$$

Indicates genetic variability at a locus



N=183

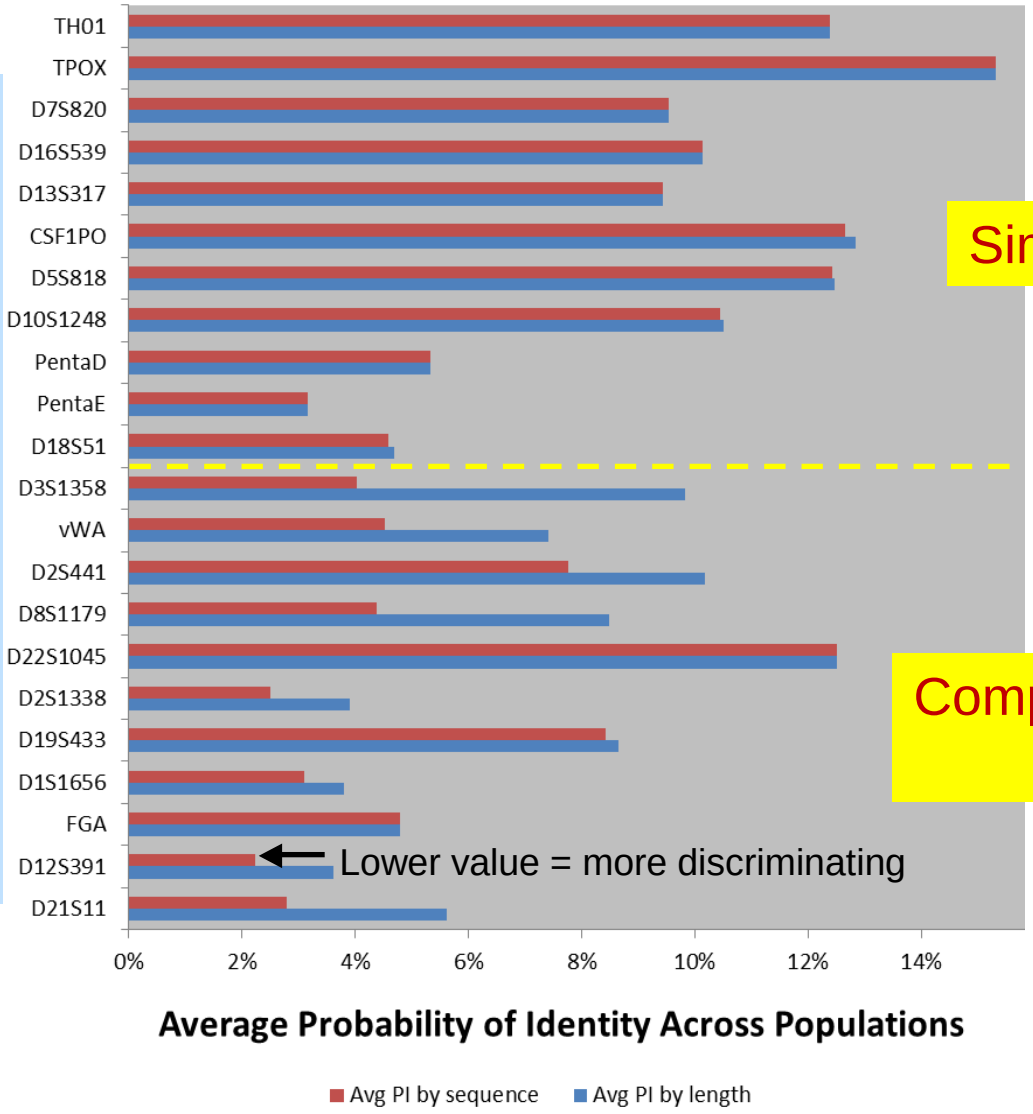
Forensic STR Sequence Diversity

Probability of Identity

Sum of each genotype frequency^2 at each locus

$= \Sigma\ xi^2$

Probability that two unrelated individuals selected at random will have the same genotype at a locus



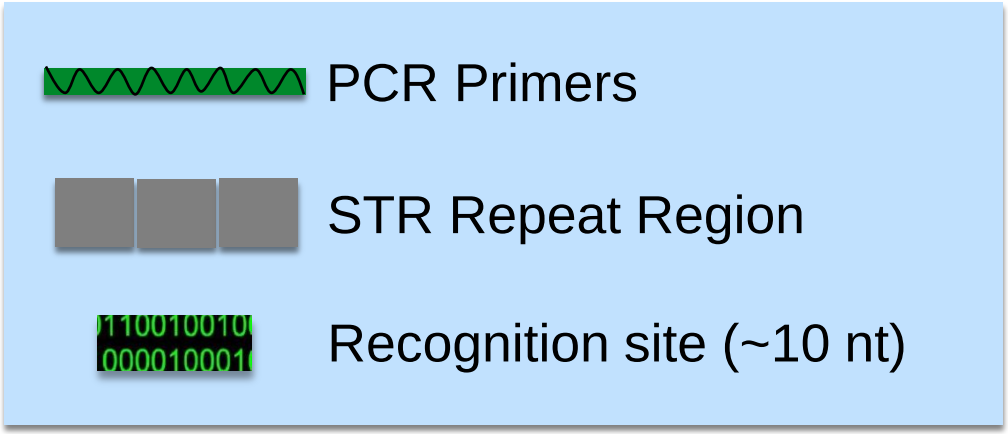
Simple Repeats

Compound/Complex Repeats

N=183

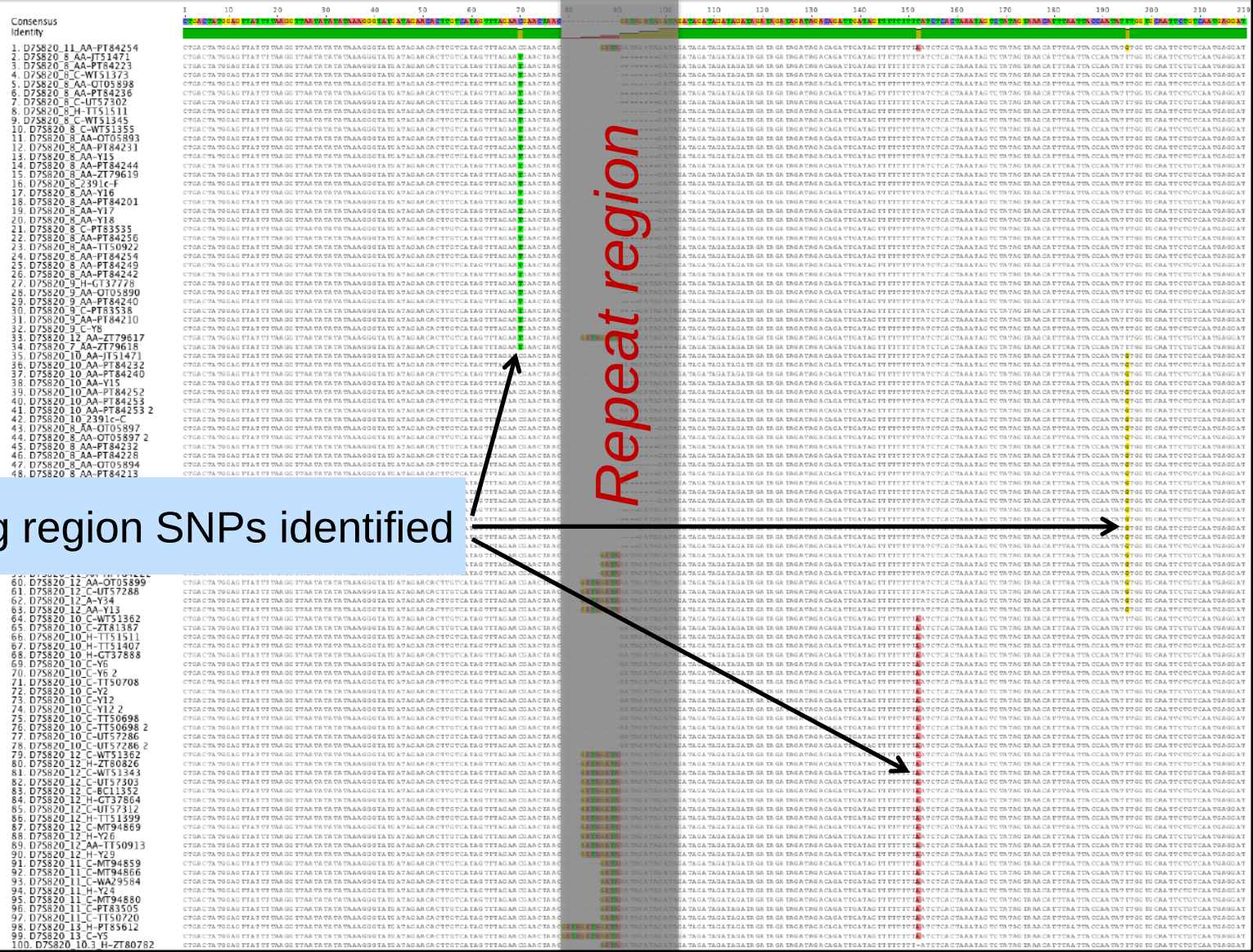
Recognition Site-Based Informatics for STRs

Moving recognition sites out will capture information within the flanking regions



Forensic STR Flanking Region

D7S820



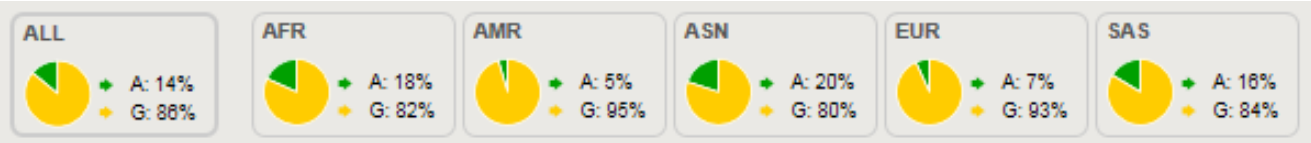
Three flanking region SNPs identified

Forensic STR Flanking Region

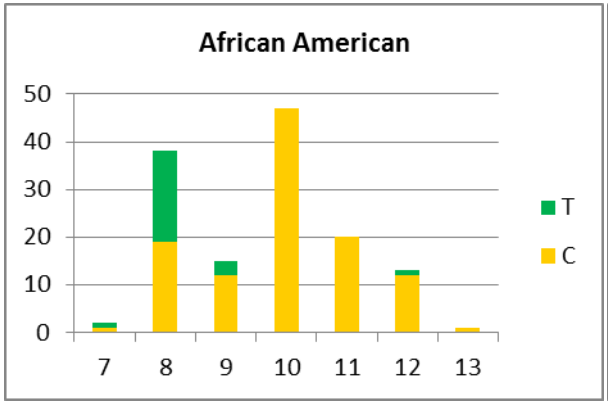
D7S820

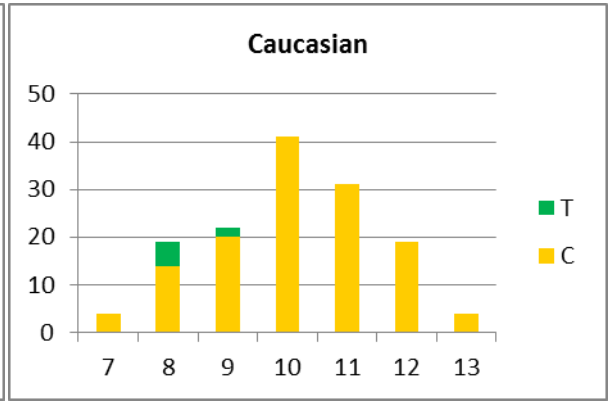
Distribution of *rs16887642* by Population in 1000 Genomes Project

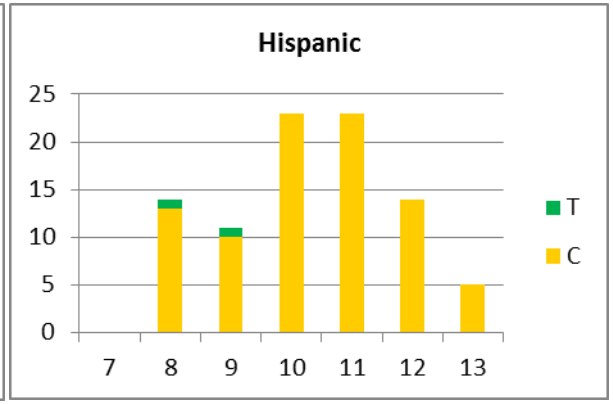
rs16887642 (note- opposite strand A/G reported):



Distribution of *rs16887642* by Population / Allele in N=183





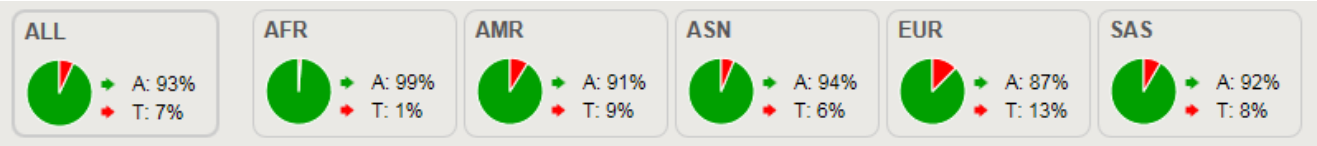


Forensic STR Flanking Region

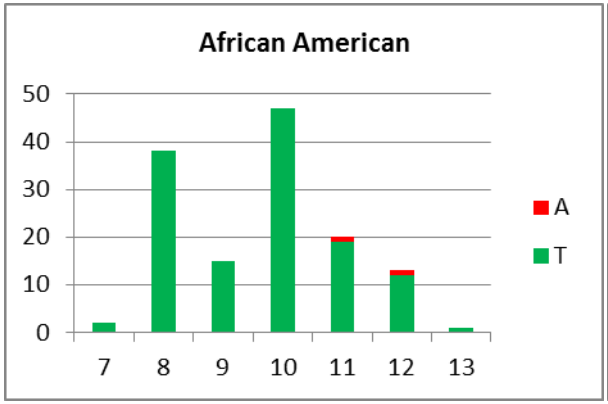
D7S820

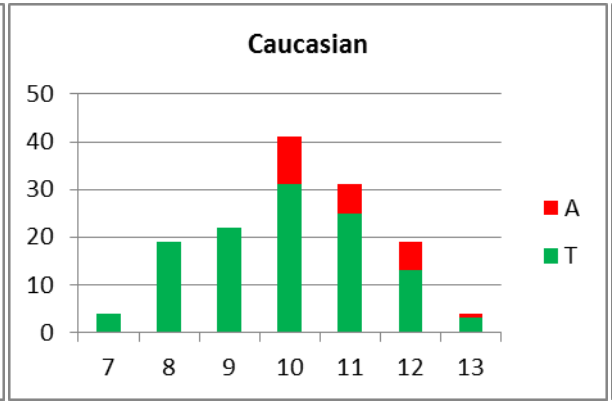
Distribution of *rs7789995* by Population in 1000 Genomes Project

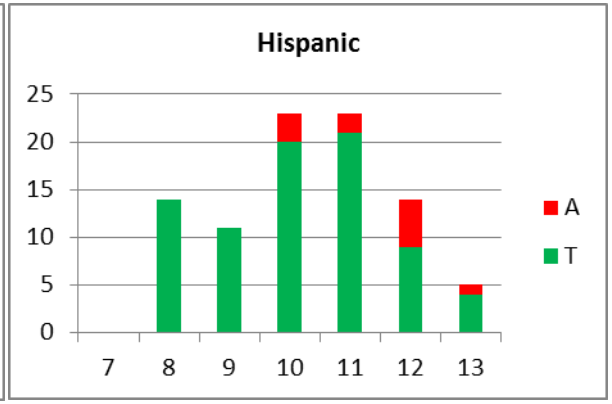
rs7789995 (note- opposite strand A/T reported):



Distribution of *rs7789995* by Population / Allele in N=183





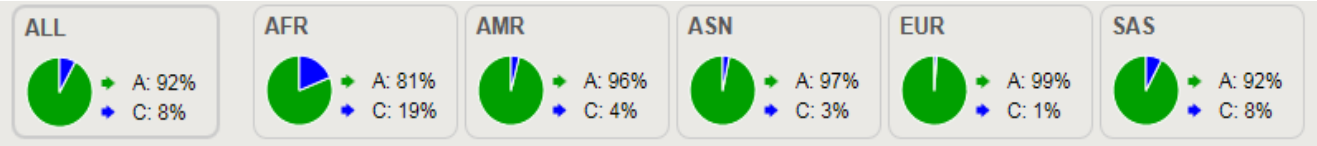


Forensic STR Flanking Region

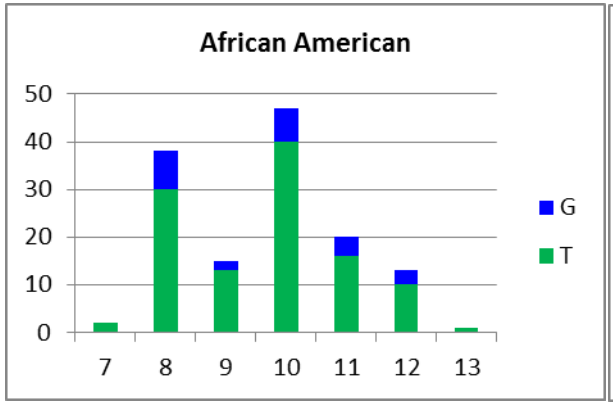
D7S820

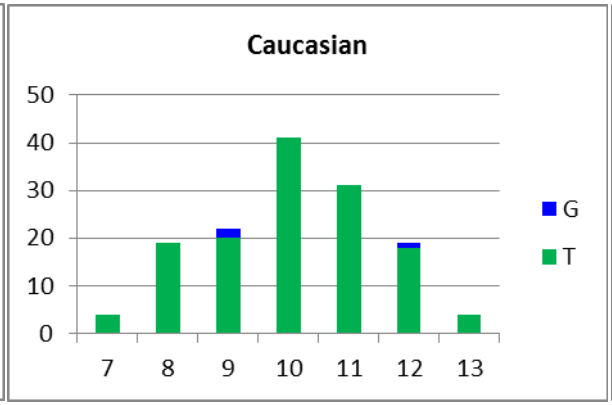
Distribution of *rs7786079* by Population in 1000 Genomes Project

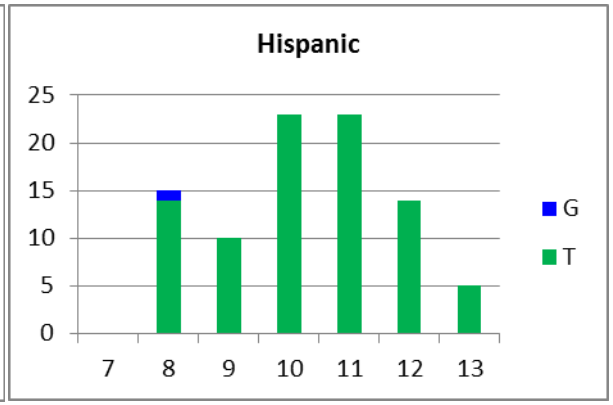
rs7786079 (note- opposite strand A/C reported):



Distribution of *rs7786079* by Population / Allele in N=183



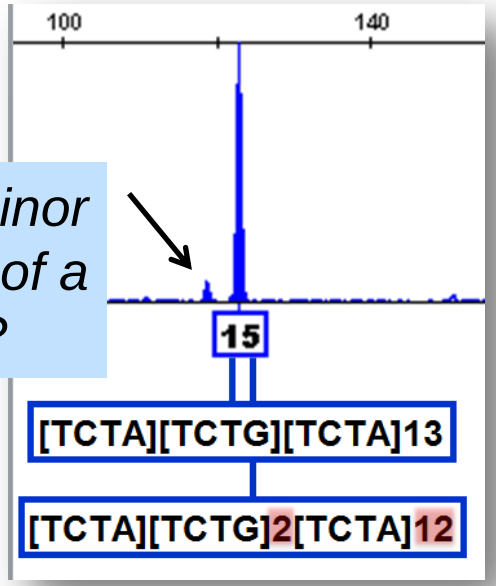




Isoalleles

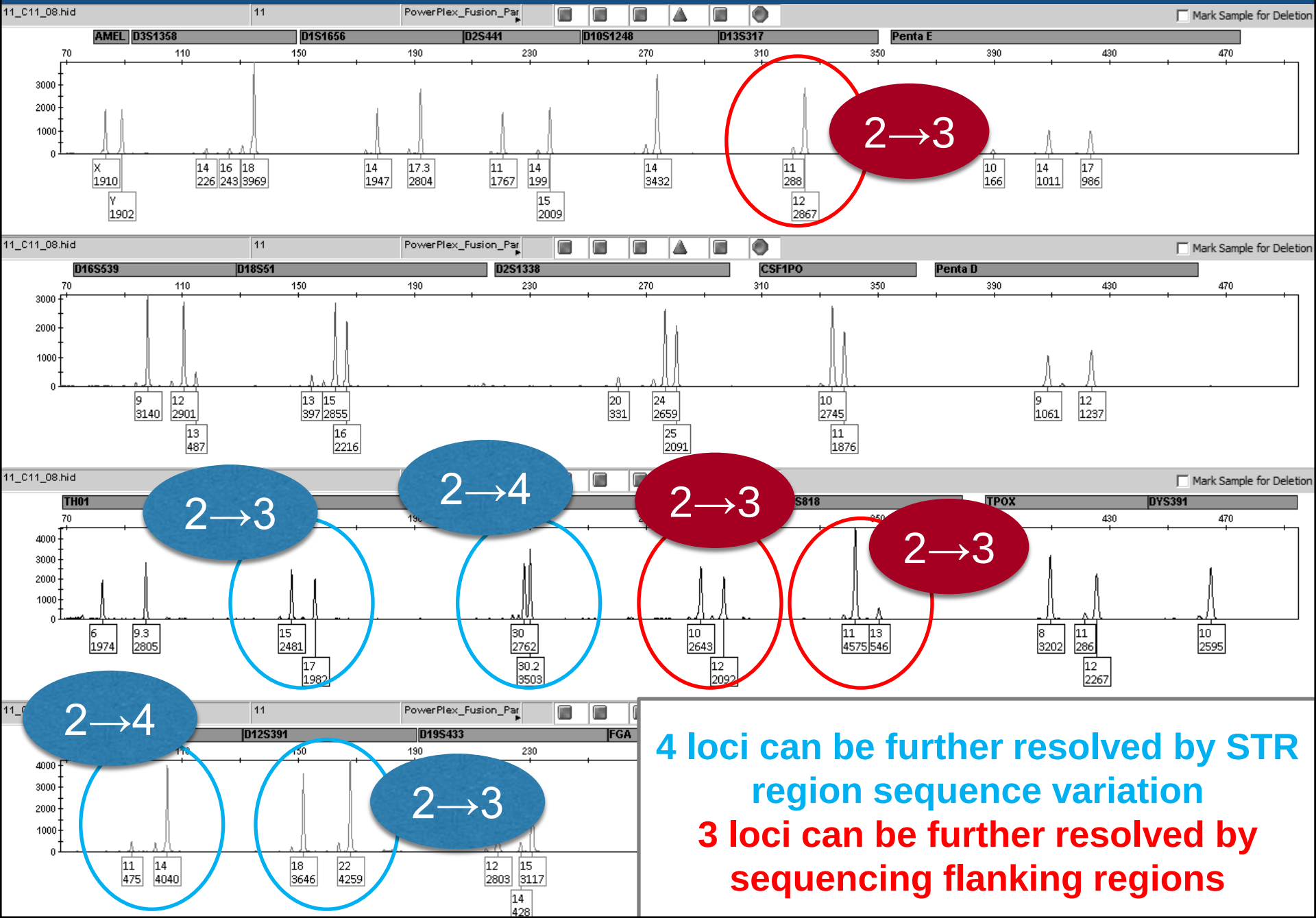
- Identical by length; unique by sequence
- Within an individual
  - Sequence based heterozygote
  - Increased heterozygosity ($p^2 \rightarrow 2pq$)
  - Marginal benefits for one-to-one matching?

Stutter or minor component of a mixture?



- Between individuals (DNA mixtures)
  - Resolve overlapping alleles 15 → ‘15A’ and ‘15B’
  - Resolve stutter from minor components

2 person, 9:1 Mixture, selected for maximal overlapping alleles



4 loci can be further resolved by STR region sequence variation

3 loci can be further resolved by sequencing flanking regions

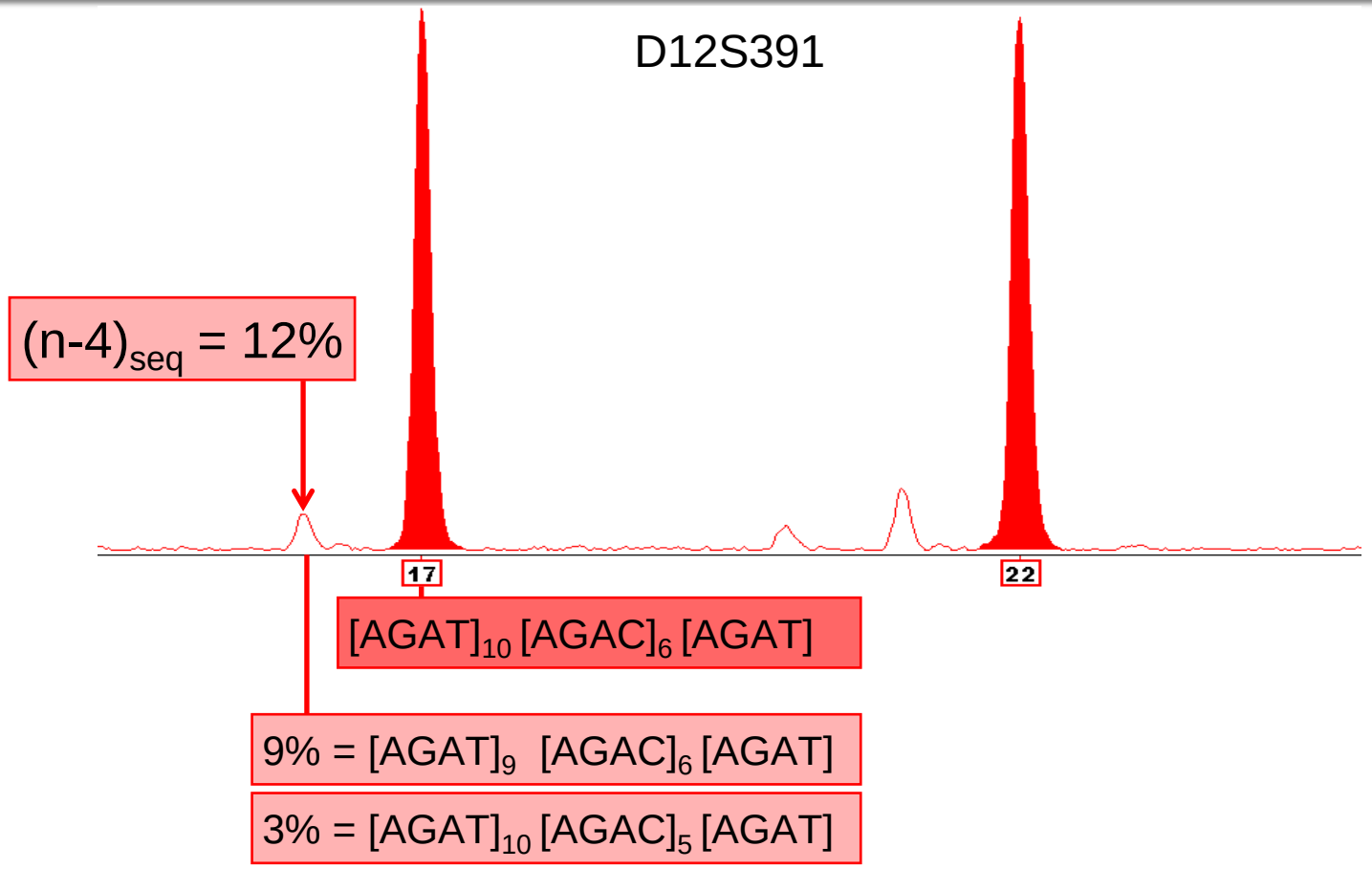
But before moving forward

In order to understand the true benefits of sequencing to mixtures

- Validate analytical and stochastic calling thresholds for NGS work
- Assess sensitivity of NGS methods to detect minor components
  - Still PCR front end – will we observe better than 10:1?
- In practice - how often will *resolvable* overlapping alleles be observed in a mixture?
- Need the allele frequencies for ‘new’ alleles
  - What size population databases are needed? Greater than 200?
- Incorporate NGS sequence data into probabilistic genotyping software (STRMix, True Allele, etc)
  - What are the gains in stats (Log LR)? Improved contributor ratio estimates?

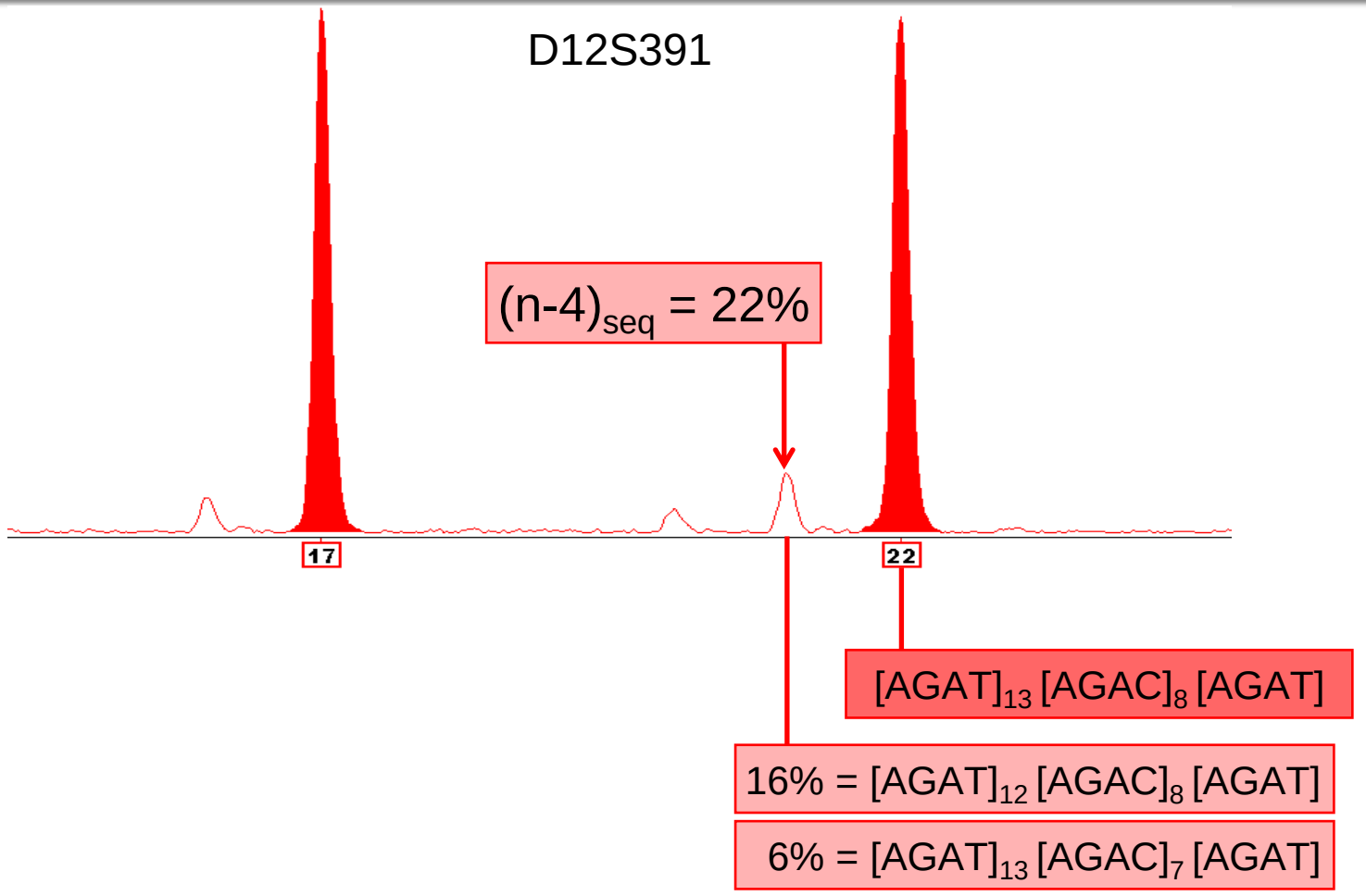
More loci – better loci?

Stutter by Sequence



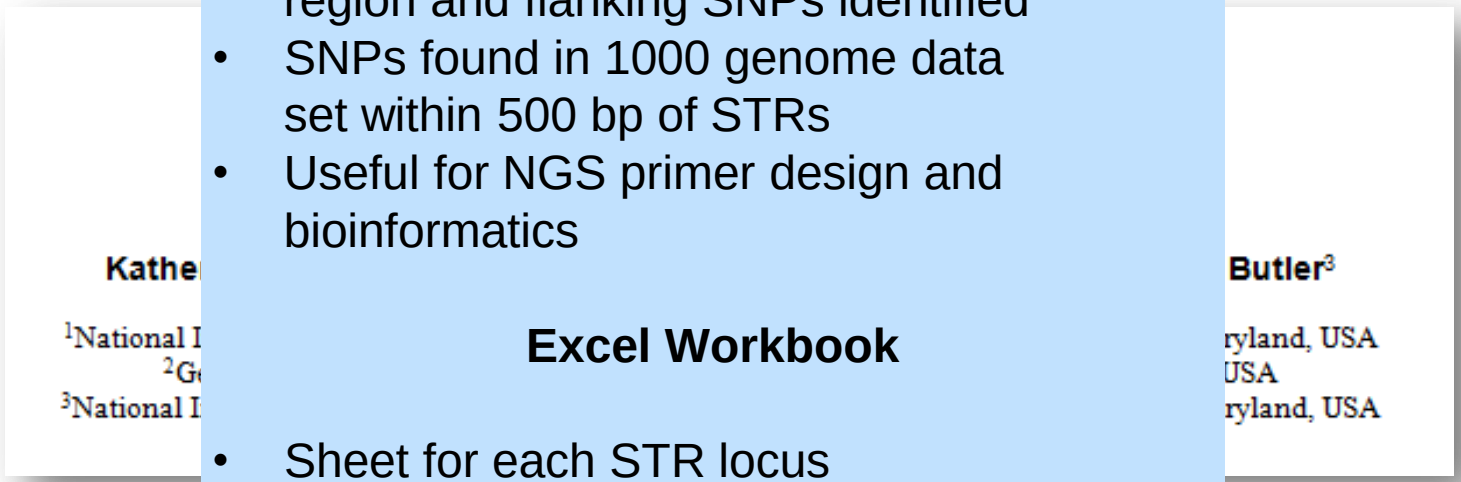
Single source sample

Stutter by Sequence



Single source sample

NIST Support for NGS Research



Annotations

- GRCh38 genome with STR repeat region and flanking SNPs identified
- SNPs found in 1000 genome data set within 500 bp of STRs
- Useful for NGS primer design and bioinformatics

Excel Workbook

- Sheet for each STR locus
- Observed alleles, repeat structure, platform
- Broken out by sub-motif
- References
- Not frequency data
- This can be updated as needed

NIST Support for NGS Research
SRM 2391c

SRM 2391c – Component B

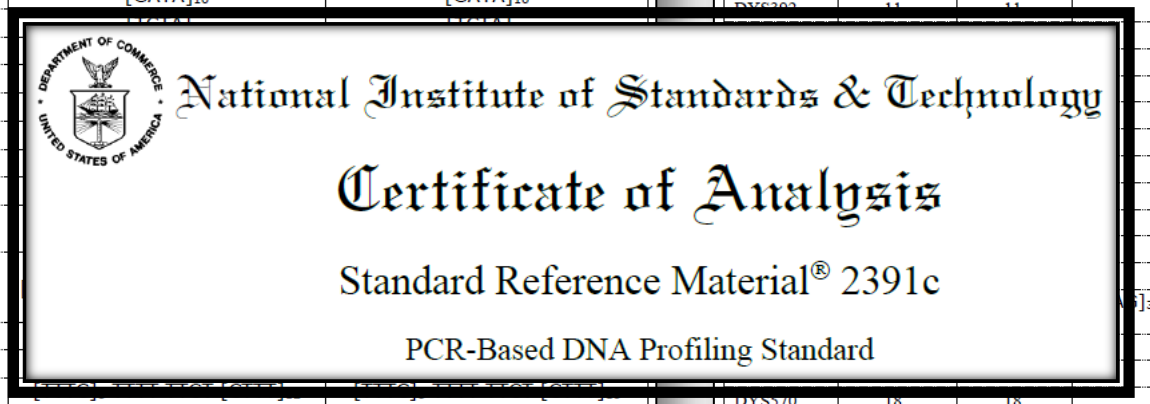
Marker	Length-based Types	Sanger Result	Repeat Structure –Allele 1	Repeat Structure –Allele 2
D1S1656	11, 14	11, 14	[TAGA]$_{11}$ [TG]$_5$	[TAGA]$_{14}$ [TG]$_5$
D2S1338	17, 17	17, 17	[TGCC]$_6$ [TTCC]$_{11}$	[TGCC]$_6$ [TTCC]$_{11}$
D2S441	10, 14	10, 14	[TCTA]$_{10}$	[TCTA]$_{11}$ TTTA [TCTA]$_2$
D3S1358	15, 19	15, 19	TCTA [TCTG]$_3$[TCTA]$_{11}$	TCTA [TCTG]$_3$[TCTA]$_{15}$
D5S818	12, 13	12, 13	[AGAT]$_{12}$	[AGAT]$_{13}$
D6S1043	14, 19	14, 19	[AGAT]$_{14}$	[AGAT]$_{13}$ ACAT [AGAT]$_5$
D7S820	10, 10	10, 10	[GATA]$_{10}$	[GATA]$_{10}$
D8S1179	10, 13	10, 13		
D8S1115	15, 17	15, 17		
D10S1248	13, 13	13, 13		
D12S391	19, 24	19, 24		
D13S317	9, 12	9, 12		
D16S539	10, 13	10, 13		
D18S51	13, 16	13, 16		
D19S433	16, 16.2	16, 16.2		
D21S11	32, 32.2	32, 32.2		
D22S1045	15, 17	15, 17		
CSF1PO	10, 11	10, 11		
FGA	20, 23	20, 23	CTCC [TTCC]$_2$	CTCC [TTCC]$_2$
Penta D	8, 12	8, 12	[AAAGA]$_8$	[AAAGA]$_{12}$
Penta E	7, 15	7, 15	[AAAGA]$_7$	[AAAGA]$_{15}$
SE33	17, 18	17, 18	[AAAG]$_2$ AG [AAAG]$_3$ AG [AAAG]$_{17}$ G [AAAG]$_3$ AG	[AAAG]$_2$ AG [AAAG]$_3$ AG [AAAG]$_{18}$ G [AAAG]$_3$ AG
TH01	6, 9.3	6, 9.3	[AATG]$_6$	[AATG]$_6$ ATG [AATG]$_3$
TPOX	8, 11	8, 11	[AATG]$_8$	[AATG]$_{11}$
vWA	17, 18	17, 18	TCTA [TCTG]$_4$ [TCTA]$_{12}$	TCTA [TCTG]$_4$ [TCTA]$_{13}$

[a] Deletion of 2 bp in an uncounted repeat unit results in the 16.2 designation.
Note: Sequence information in gray indicates bases that are not counted toward the length-based genotype designation.

SRM 2391c – Component B

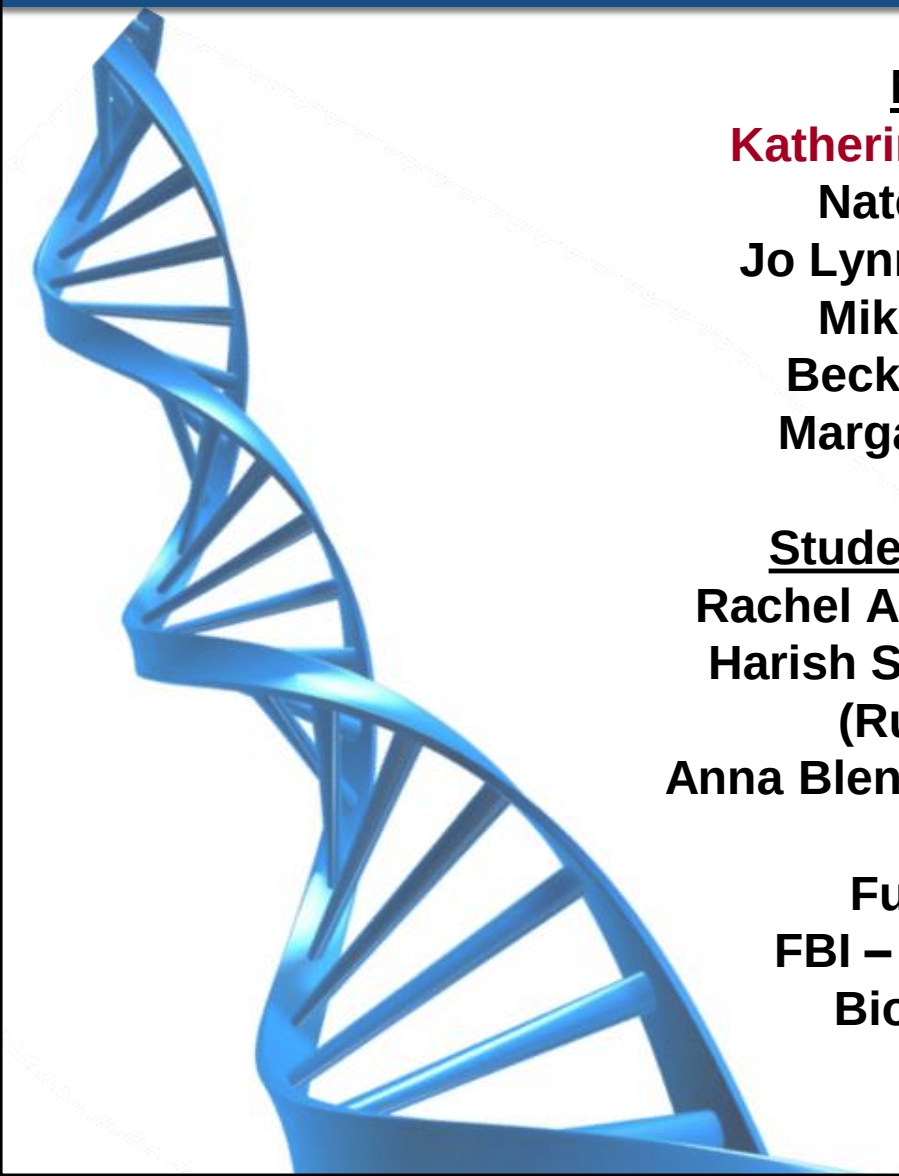
Marker	Length-based Types	Sanger Result	Repeat Structure –Allele 1
DYS19	14	14	[TAGA]$_3$ TAGG [TAGA]$_{11}$
DYS385a	13	13	[GAAA]$_{13}$
DYS385b	17	17	[GAAA]$_{17}$
DYS389I	13	13	[TCTG]$_3$ [TCTA]$_{10}$
DYS389II	31	31	[TCTG]$_6$ [TCTA]$_{12}$ N$_{48}$ [TCTG]$_3$ [TCTA]$_{10}$
DYS390	23	23	[TCTG]$_8$ [TCTA]$_{10}$ TCTG [TCTA]$_4$
DYS391	10	10	[TCTA]$_{10}$
			[TAT]$_{11}$
			[AGAT]$_{12}$
			[TCTA]$_8$ [TCTG]$_2$ [TCTA]$_4$
			[TTTTC]$_{10}$
			[AGAT]$_{11}$
			[AGAGAT]$_{12}$ N$_{42}$ [AGAGAT]$_8$
			[TTTC]$_{11}$ N$_{50}$ [TTTC]$_{15}$
			[AGAT]$_{15}$
			[GAAA]$_{15}$ AA [GAAA]$_2$
			[ATAG]$_{10}$
			[CTT]$_{25}$
			]$_3$ GAAG [AAAG]$_{14}$ GGAG [AAAG]$_4$ N$_6$ [AAAG]$_{15}$
			[ATCT]$_{11}$
			[GATA]$_{12}$
DYS570	18	18	[TTTC]$_{18}$
DYS576	17	17	[AAAG]$_{17}$
DYS627	22	22	[AGAG]$_3$[AAAG]$_{19}$
DYS635	20	20	[TCTA]$_4$ [TGTA]$_2$ [TCTA]$_2$ [TGTC]$_2$ [TCTA]$_{10}$
DYS643	9	9	[CTTTT]$_9$
DYF387S1a	35	35	[AAAG]$_3$ GTAG [GAAG]$_4$ N$_{20}$ [GAAG]$_9$ [AAAG]$_{13}$
DYF387S1b	38	38	[AAAG]$_3$ GTAG [GAAG]$_4$ N$_{20}$ [GAAG]$_{10}$ [AAAG]$_{15}$
Y GATA H4	11	11	[TAGA]$_{11}$

Note: Sequence information in gray indicates bases that are not counted toward the length-based genotype designation.



http://www.nist.gov/srm/

Acknowledgements



NIST

Katherine Gettings
Nate Olson
Jo Lynne Harenza
Mike Coble
Becky Steffen
Margaret Kline

Student Interns

Rachel Aponte (GWU)
Harish Swaminathan (Rutgers)
Anna Blendermann (MC)

Funding

FBI – DNA as a Biometric

Battelle

Seth Faith (now @NCSU)
Rich Guerrieri
Brian Young
Liz Montano
Esley Heizer
Angela Minard-Smith
Christine Baker

Promega

Doug Storts
Jay Patel

Contact Information
peter.vallone@nist.gov